Health Care Manag Sci (2014) 17:88–99
DOI 10.1007/s10729-013-9241-3

Estimating the waiting time of multi-priority emergency patients with downstream blocking

Di Lin · Jonathan Patrick · Fabrice Labeau

Received: 12 November 2012 / Accepted: 22 April 2013 / Published online: 21 May 2013

Abstract To characterize the coupling effect between patient flow to access the emergency department (ED) and that to access the inpatient unit (IU), we develop a model with two connected queues: one upstream queue for the patient flow to access the ED and one downstream queue for the patient flow to access the IU. Building on this patient flow model, we employ queueing theory to estimate the average waiting time across patients. Using priority specific wait time targets, we further estimate the necessary number of ED and IU resources. Finally, we investigate how an alternative way of accessing ED (Fast Track) impacts the average waiting time of patients as well as the necessary number of ED/IU resources. This model as well as the analysis on patient flow can help the designer or manager of a hospital make decisions on the allocation of ED/IU resources in a hospital.

Keywords Emergency department · Waiting time · Queueing theory · Hospital management

1 Introduction

Overcrowding in the emergency department (ED) is a worldwide problem [1–3] impairing the ability of hospitals to offer emergency care within a reasonable time frame [4]. By observing more than 20 million patient visits to emergency departments over five years, Guttmann et al. in [5] determined that the risk of death and hospital readmission increases with the degree of crowding in the emergency department, and estimated that about 150 fewer patients would die in Ontario each year if the average waiting time to access the emergency department was less than an hour.

The issue of what constitutes timely access to emergency care is obviously dependent on the acuity of the patient. Without an accurate triage and acuity scale, patients who need immediate emergency care will experience a delay in treatment that may aggravate their condition. The Canadian government published its own acuity guidelines in 1998, and subsequently revised them in 2004 and in 2008. In these guidelines, the severity of patients is classified into five levels: resuscitation, emergent, urgent, less urgent and non urgent [6]. This classification is based on a patient's presenting complaints, vital signs (including the hemodynamic stability, hypertension, temperature, level of consciousness, respiratory distress, etc.), pain severity, and injury level. For patients in each severity level, their target waiting time to see a physician is detailed in Table 1.

However, strict adherence to the priority system may mean that a low priority patient may wait a long time to receive a relatively simple procedure that would tie up ED resources for very little time whereas the higher acuity patients have complex needs that require significant resources. As a means of addressing this issue, many hospitals have introduced a fast-track system for the lower (less urgent and non-urgent) priority patients on the premise that they can be served quickly and easily without tying up too many resources. This policy essentially means that a single queue is broken into two and the ones who inevitably suffer from such a policy are the patients at the end of the first queue (priority III patients). Thus, though the impact of

D. Lin (✉) · F. Labeau
Department of Electrical and Computer Engineering,
McGill University, McConnell, 633 3480 University Street,
Montreal, Quebec, Canada
e-mail: di.lin2@mail.mcgill.ca

J. Patrick
Telfer School of Management, University of Ottawa, DMS 7151,
55 Laurier Avenue East, Ottawa, ON K1N 6N5, Canada

Table 1 Triage levels for emergency department [9]

Triage level	Expected waiting time to see a physician
I: Resuscitation	Immediate
II: Emergent	<15 min
III: Urgent	<30 min
IV: Less urgent	<60 min
V: Non urgent	<120 min

the fast-track system is to reduce over-crowding by serving low acuity patients quickly, it is our contention that, unless resources are increased, this is accomplished at the cost of increasing the wait times for priority III patients. This contention has been partially confirmed. Cooke et al. in [7] and Miquel et al. in [8] show that the fast track can reduce the waiting time for patients who are qualified to access the fast track but will slightly lengthen the waiting time of the other patients. In this paper, we therefore concentrate on a standard priority system without a fast-track but also provide some results for a system with a fast track.

Even with a good acuity scale in place, insufficient resources can still cause overcrowding. As shown in Fig. 1, the resources available in the ED and in the inpatient unit (IU) will influence patient flow potentially leading to long wait times to access the ED. While the lack of ED resources can block patients in the waiting room, an insufficient number of IU resources may block transfers to the IU further delaying other patients in the waiting room from accessing the ED. Limited budgets however mean that it is inefficient to carry too many resources either in the ED or the IU. Thus, it is imperative that a methodology be developed that correctly estimates the necessary resource capacity in a hospital in order to provide timely access to the ED both to avoid overcrowding and idle time. In this paper, we concentrate on bed capacity in the ED and the IUs. Clearly there are other resources that play a role in determining the service time in the ED such as lab capacity and testing equipment. The assumption here is that other resources are not the rate limiting step to meeting the targets set out in Table 1.

In this paper, we develop a queueing model to estimate the waiting time of patients to access the ED as well as the necessary amount of resources to achieve the wait time targets for each priority class. This queueing model is composed of two queues: the first one is a $M/G/c_1/\infty$ with five priorities. In the above notation, the M refers to the assumption that the arrival of demand follows a Poisson distribution, the G means that the service rate can follow a general distribution and the c_1 refers to the number of available resources in the ED. This queue characterizes the patient flow from accessing the ED to departing the ED. The five priorities represent five severity levels classified by Table 1. The second queue is a $G/GI/c_2/c_2$ (general arrival distribution, general but independent service times with c_2 servers and c_2 capacity in IU) without priority and no buffer, and it characterizes the patient flow from accessing the IU to discharge.

Additionally, in order to represent the coupling effect that occurs when patients are blocked from entering the IU, we estimate the probability of all servers being busy in the second queue and the probability of patients in the ED waiting for transfer to the IU being blocked. Building on this queueing model, we attempt to estimate the necessary number of ED and IU resources to achieve the performance targets of Canadian Triage And Acuity Scale (CTAS), shown in Table 1.

2 Related work

Studies relevant to the estimation of the necessary amount of ED resources can be classified into two stages: (1) Rough estimation stage and (2) further adjustment stage. First, the ED manager needs to estimate the average resource requirements over an extended period of time. Studies in this stream usually estimate the number of resources by the steady state of their estimation models providing a

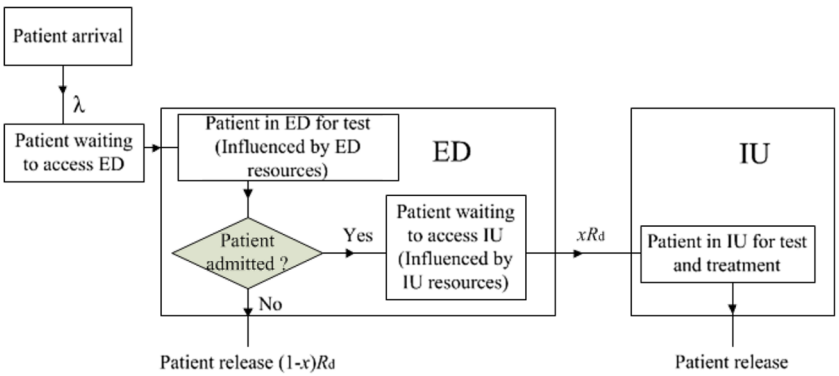

Fig. 1 Patient flow to access and depart ED [1]

long run average estimation of the resource requirements. This would constitute the first stage. Afterwards, the ED manager may need to adjust the resource assignment to meet the daily fluctuations in demand that are characteristic of ED departments [10–12].

From the perspective of methodologies, while the studies of the first type often utilize queueing [2, 10–12] and Markov chain models [3] to estimate the necessary amount of ED resources in the steady state, the studies of the second type usually utilize an autoregressive integrated moving average model (ARIMA) [13–15], Monte-Carlo simulation [1, 17], or Markov decision process (MDP) [16, 19, 20] to dynamically analyze the necessary amount of resource as a function of time. From the perspective of the length of the planning horizon, the adjustment stage studies can further be classified into short-term estimation and long-term estimation. For short-term estimation (the length of the planning horizon is at most one day) the variation of arrival rate is measured at the hourly level [13–16]. For long-term estimation (the length of observation ranges from more than one day to a few months), the daily variation or even seasonal variation is factored into the planning decisions [19, 20].

Our study is of the first type, namely, estimating the steady state resource requirements in a hospital in order to provide timely access to the ED. To our knowledge no paper has dealt with both the coupling effect between two resources (such as the ED and IU) as well as the presence of multiple priority classes in the demand stream. Patrick et al. in [19, 20] present how to estimate the required capacity for a diagnostic imaging department dealing with multiple priority classes but deal with only a single resource. Kolb et al. in [1, 17] take into account the coupling effect between the ED and the IU but do not consider the different triage levels of patients. Similarly, Koizumi et al. in [21] present a model that deals with a series of resources with potential blocking but with no acuity scale differentiating patients. For the purposes of estimating resource requirements in the ED to achieve priority-specific performance targets, it is important to propose a model that takes into account both factors. The model proposed in the next sections achieves that goal.

3 Models of patient flow

In this section, we present the arrival process of emergency patients as well as the model to describe this process. Building on this model, we estimate the average waiting time of an emergency patient accessing the ED.

Patients arrive at the ED either as walk-in patients or ambulatory patients. After entering the main entrance, the patient moves from the greeter desk to triage to registration before waiting for a physician consult. In our model these steps are combined under the heading 'Patients waiting to access the ED' (see Fig. 1). Once a patient accesses the ED, they are assigned a bed by a nurse. After a pre-examination by the ED nurse, the patient will see a physician for further examination, tests and treatments. If the physician determines the patient can be discharged after examination, an ED nurse facilitates the process of releasing the patient who then departs the ED. If the physician decides to transfer the patient to the inpatient units (IUs), a consulting physician from the IU will arrange the admission. Depending on the availability of IU beds, the patient is either transferred immediately or has to wait in the ED till an IU bed becomes available. These steps in our model are combined under the heading 'Patients staying in the ED'. A patient in the IU will receive tests and treatments until he/she is deemed ready for discharge. After a final examination by a physician, the patient will depart the IU. These steps in our model are combined under the heading 'Patients staying in the IU'.

3.1 Arrival process of demand accessing the ED and the IU

Demand for ED service fluctuates significantly throughout the day. In consequence, Koizumi et al. in [21] divide the whole day into several segments such that the arrival rate in each segment can be assumed to be constant. They demonstrate that the arrival process of emergency patients in most time segments can be modeled as a Poisson process with varying rates.

As mentioned in the previous section, our model looks to provide the steady state resource requirements over a longer period than a day and thus we do not take into account the within day variations. Given the results from our model, a further model (perhaps using a Markov Decision Process approach) could be used to allocate the available resources in order to take into account the within-day fluctuations in demand.

The data presented in Table 2 was provided by a local hospital and represents the average arrival rate of patients entering the ED per hour (broken down by CTAS) and the

Table 2 Patient flow into the ED and the IU

Destination	CTAS/Point of Entry	Arrival Rates
ED	I	0.075
	II	0.662
	III	3.749
	IV	2.86
	V	0.226
	All	7.572
IU	From ED	0.479
	Direct	0.267
	All	0.746

average arrival of patients entering the IU per hour (broken down by point of entry) for fiscal year 2011/2012.

3.2 Queueing models to estimate the waiting time

Building on the flowchart of patients accessing and departing the ED (shown in Fig. 1), we use queueing theory to estimate the waiting time of a patient to access the ED with the waiting time of a patient referring to the time between arriving at the waiting room until treatment. To estimate the waiting time, we need to model two patient flows. (1) The patient flow from arriving at the waiting room until departing the ED can be viewed as a $M/GI/c_1/\infty$ queue to model the patient flow arriving at and departing from waiting rooms or ambulances since the size of the waiting room is rarely the limiting factor. A $M/GI/c_1/\infty$ queue has a Poisson arrival process, independent and identically distributed service times with a general distribution, c_1 ED beds and infinite buffer capacity. (2) The patient flow into the IU through to discharge can be viewed as a $G/GI/c_2/c_2$ queue, which has a general arrival process, independent and identically distributed service times with a general distribution, c_2 IU beds and no buffer capacity.

Model for patient flow in the ED Let λ represent the arrival rate of emergency patients and μ_1 represent the average service rate in the ED without taking into account the coupling effect. Due to the potential for patients being blocked from leaving the ED, we can adjust the average service time in the ED, $1/\mu$, (represented by a $M/G/c_1/\infty$ queue) as

$$1/\mu = 1/\mu_1 + P_b \times mean(\min_i T_i) \tag{1}$$

where T_i $(i = 1, 2 \cdots c_2)$ is the waiting time until ith inpatient is discharged and P_b is the blocking probability to access the IUs. The rate of patients transferring to the IU is xR_d, where $R_d = min\{\lambda, c_1\mu\}$ and x is the proportion of patients transferred into the IU.

Thus, this modified average service time represents the actual service in the ED plus the time spent waiting in the ED for an IU bed. The time waiting to access the IU equals the average length of waiting until the next inpatient departure, denoted as $\min_i T_i$.

To incorporate the priority classification system of emergency patients, we use the theory for a preemptive resume multi-priority $M/G/c_1/\infty$ queue to calculate the length of waiting time to access the ED. The rationale for using preemptive resume queue models is that high priority (Level I or level II) patients must receive immediate service. If no physician is available when such a patient arrives, a physician treating a lower priority patient (or stabilized higher priority one) must leave his/her current patient and resume only after they offer the emergency treatment to the high priority patient. The detailed calculation on the length of the waiting time is presented in Section 4.

Model for patient flow in the IU The flow of patients accessing and departing the IUs can be modeled as a $G/GI/c_2/c_2$ queue [22]. In this queueing model, we focus on the blocking probability P_b, that is, the probability that there are no available beds in the IUs. This can be determined by

$$P_b = \frac{\alpha\beta e^{-k\beta/v}}{(1 - e^{-k\beta/v})\rho_d\sqrt{c_2}} \tag{2}$$

where c_2 is the number of beds available in the IUs, λ_d is the arrival rate of patients directly accessing the IUs and μ is the service rate. Other parameters in Eq. 2 can be calculated as:

$$\rho_d = \frac{xR_d + \lambda_d}{c_2\mu}, \qquad \beta = \sqrt{c_2}(1 - \rho_d),$$

$$k = \sqrt{c_2}, \qquad v = \frac{1 + C_a^2}{2},$$

$$\alpha = \left[1 + \beta\Phi(\beta)\big/\varphi(\beta)\right]^{-1},$$

where C_a^2 is the squared coefficient of variation (SCV) of the service time at inpatient units and $\Phi(\cdot)$ and $\varphi(\cdot)$ are the cumulative distribution function (CDF) and probability density function (PDF) of a standard normal distribution.

3.3 Steady state conditions of queuing models

In this paper, we employ queueing models to represent the steady state of a queueing system to access the ED. As shown in Fig. 1, the patient flow for accessing and departing the ED is represented by two coupled queueing models. In the following, we will explore the conditions for both queueing models to reach the steady state.

Given the arrival rate λ, the necessary and sufficient steady state condition for the first queue is $\lambda \leq c_1\mu$, and that for the second queue is $xR_d + \lambda_d \leq c_2\mu_I$. Given the parameters of these two queues, we can decide whether these queues can reach steady state if we know the exact μ, which determines both the service rate in the first queue and the arrival rate of the second queue. Unfortunately μ is interdependent with another unknown parameter P_b (shown in Eqs. 1 and 2), and neither of them can be shown in a closed form. Thus, the key to attaining the necessary and sufficient steady state conditions for both queues is to calculate μ. In the following, relaxing the constraints, we first present a sufficient (not necessary) condition as well as a necessary (not sufficient) condition without requiring the exact μ. Second, we develop an iterative algorithm to numerically compute μ from which a necessary and sufficient steady state condition can be approximated.

Sufficient (but not necessary) and necessary (but not sufficient) conditions As the blocking probability satisfies $0 \leq P_b \leq 1$, we can obtain constraints on μ as $\mu_{min} \leq \mu \leq \mu_{max}$, where

$$\mu_{min} = 1 \Big/ (1/\mu_1 + mean(\min_i T_i)) \text{ and} \tag{3}$$

$$\mu_{max} = \mu_1, \tag{4}$$

where μ_{min} is the value of μ (refer to Eq. 1) when $P_b = 1$ (the worst case in which an ED patient transferring into IU is always blocked), and μ_{max} is the value of μ when $P_b = 0$ (the best case in which an ED patient transferring into IU is never blocked). Thus, the sufficient steady state condition of the first queue is $\lambda \leq c_1\mu_{\min}$, and its necessary steady state condition is $\lambda \leq c_1\mu_{\max}$. Correspondingly, the sufficient steady state condition of the second queue is $x \min\{\lambda, c_1\mu_{\min}\} + \lambda_d \leq c_2\mu_I$, and its necessary steady state condition is $x \min\{\lambda, c_1\mu_{\max}\} + \lambda_d \leq c_2\mu_I$.

Necessary and sufficient condition In the following, we will discuss the necessary and sufficient steady state conditon. First of all, we represent Eq. 1 as

$$y_1(\mu) + y_2(\mu) - 1/\mu = y_0 \tag{5}$$

where $y_1(\mu) = 1/\mu$, $y_2(\mu) = 1/\mu - P_b \times mean(\min_i T_i)$, and $y_0 = 1/\mu_1$. In the following, we show that there is at least one feasible solution to Eq. 5, and this solution $\hat{\mu} \in [\mu_{min}, \mu_{max}]$.

Proof We will prove it by contradiction. Let $y(\mu) = y_1(\mu)+y_2(\mu)-1/\mu$. From Eq. 2, we know $y_2(\mu) \leq y(\mu) \leq y_1(\mu)$. Assume that there is no feasible solution to Eq. 5, then $y_0 = y_2(\mu) < y(\mu)$ at $\mu = \mu_{min}$, otherwise μ_{min} will be the feasible solution. Because of the continuity of $y(\mu)$ and our assumption of no feasible solution in $[\mu_{min}, \mu_{max}]$, we can deduce that $y(\mu) > y_0$ for $\mu \in [\mu_{min}, \mu_{max}]$. At $\mu = \mu_{max}$, we can conclude that $y(\mu) > y_0 = y_1(\mu)$, which is contradicted by the fact $y(\mu) \leq y_1(\mu)$ from Eq. 2. □

Beyond the proof, we can also demonstrate that at least one solution exists intuitively from Fig. 2. Certainly a feasible solution exists (shown in circle), because of the constraints: $y_2(\mu) \leq y(\mu) \leq y_1(\mu)$, $y_2(\mu_{min}) = y_0$, and $y_1(\mu_{max}) = y_0$.

If there is only a single solution $\hat{\mu}$, then it is easy to show the necessary and sufficient steady state condition as $\lambda \leq c_1\hat{\mu}$ for the first queue and $x \min\{\lambda, c_1\hat{\mu}\} + \lambda_d \leq c_2\mu_I$ for the second queue.If there are multiple solutions, then no necessary and sufficient steady state condition exists. In the latter case, we can at least find a tighter sufficient condition and a tighter necessary condition by replacing μ_{min} and μ_{max} by $\hat{\mu}_{min}$ and $\hat{\mu}_{max}$, respectively. Please note that $\hat{\mu}_{min}$ and $\hat{\mu}_{max}$ are the minimal and maximal feasible solutions. The detailed process of searching for the minimal and maximal feasible solutions is shown in Algorithm 1.

Algorithm 1 Numerical estimation of μ

Input: Equation (5)
Output: Return the minimal and maximal feasible solutions

1 Search the minimal feasible solution $\hat{\mu}_{min}$ by starting with μ_{min};
2 Search the maximal feasible solution $\hat{\mu}_{max}$ by starting with μ_{max};
3 If $\hat{\mu}_{max} = \hat{\mu}_{min}$, output 'A single solution' and $\hat{\mu}_{max}$;
4 If $\hat{\mu}_{max} \neq \hat{\mu}_{min}$, output 'Multiple solutions' and $\hat{\mu}_{max}$ as well as $\hat{\mu}_{min}$;

If $\hat{\mu}_{min}$ and $\hat{\mu}_{max}$ [1] are the same, the sufficient and necessary condition exists and this condition is: $\lambda \leq c_1\hat{\mu}$ for the first queue and $x \min\{\lambda, c_1\hat{\mu}\} + \lambda_d \leq c_2\mu_I$ for the second queue. Otherwise, there is no sufficient and necessary condition, and we can use the tightest sufficient condition: $\lambda \leq c_1\hat{\mu}_{min}$ for the first queue and $x \min\{\lambda, c_1\hat{\mu}_{min}\}+\lambda_d \leq c_2\mu_I$ for the second queue to guarantee the stability of the queue.

4 Estimation of the waiting time using the queueing models

In this section, we present the estimation of the waiting time in the ED. As shown in Eq. 1, this waiting time is dependent on $mean(\min_i T_i)$, namely, the average waiting time to be transferred to the IUs. In the following, we first estimate $mean(\min_i T_i)$ and then estimate the waiting time in the ED. Also we take into account two cases: patient flow with a fast track and without a fast track. Specifically, we present the estimation of waiting time in the case of no fast track in IV. B., and the case with a fast track in IV. C.

4.1 Estimation of the waiting time to be transferred into the IUs

In this section, we discuss how to calculate the average waiting time to access the IUs, namely, $mean(\min_i T_i)$. Given $F_i(z)$ as the cumulative distribution function (CDF) of T_i $(i = 1, \cdots, c_2)$, which is estimated as a triangular distribution in [17], the CDF of $\min_i T_i$ (denoted as $F_{min}(z)$)

[1] Searching for the closest solution from the starting point can be done by any standard method. We use the Matlab function 'fsolve' to carry out the search, and its default internal algorithm is the trust-region dogleg algorithm [23].

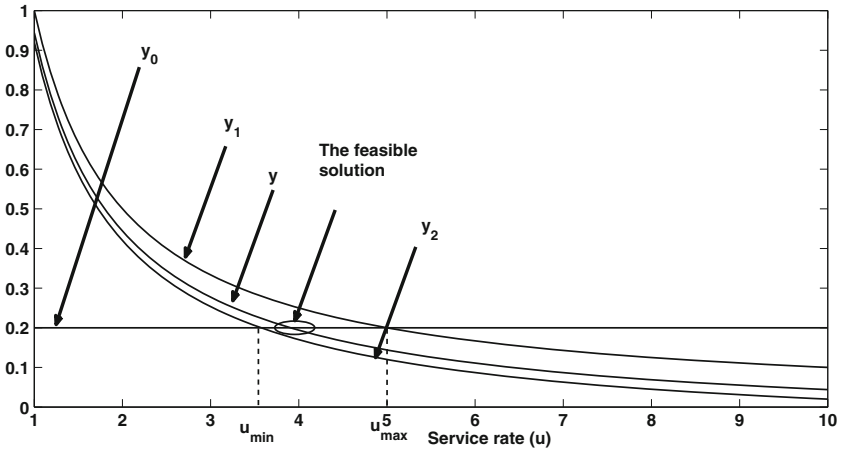

Fig. 2 Intuitively understanding the proof

can be expressed as [18]

$$F_{\min}(z) = 1 - \prod_{i=1}^{c_2} [1 - F_i(z)] \tag{6}$$

Building on Eq. 6, we can calculate the mean of $\min_i T_i$ as

$$mean(\min_i T_i) = \int z dF_{\min}(z) \tag{7}$$

By substituting Eqs. 6 and 7 into Eq. 1, we can attain the expression of the modified average service time $1/\mu$.

4.2 Estimation of the waiting time in the ED

As presented in Table 2, emergency care patients can be classified into 5 categories. Patients with severity level I and II are provided priority over ambulatory patients waiting for admittance to the ER. The prioritizing frequently includes pre-empting or suspending service for patients fitting the profile of categories IV and V so that appropriate care can be provided in life threatening situations [29]. From the perspective of queuing models, a pre-empting multiple-priority queue is appropriate for this scenario [28–30].

As presented in III.B.(1), we will use a $M/G/c/\infty$ queue to characterize the patient flow. Few studies provide the exact closed-form expression of waiting time in a $M/G/c/\infty$ queue. The exact waiting time is difficult to calculate because such a queue does not possess an embedded Markov chain [24]. Instead, most studies focus on an approximation of a $M/G/c/\infty$ queue. Hokstad in [25] and Miyazawa in [26] both suggest using a probability generating function (PGF) to generate an approximation of $M/G/c/\infty$ queues. Tijms et al. in [27] propose a regenerative approach to simulate a $M/G/c/\infty$ queue. The aforementioned approximation methods are reliable but all require knowledge of the exact distribution of the service time, which is not applicable in our scenario because the distribution of our modified service time is unknown (shown in Eq. 1).

By using heuristics, Bondi et al. in [31] present the estimation of the waiting time in a preemptive resume multi-priority $M/G/c_1/\infty$ queue. Given the number of priorities K, the mean waiting time of patients with priority k ($k = 1, \cdots, K$), $W_{c_1}^k$, can be estimated as [31]

$$W_{c_1}^k \approx \frac{W_{c_1}^F W_1^k}{W_1^F} \tag{8}$$

where W_1^k represents the mean waiting time of patients with priority k in a preemptive resume multi-priority $M/G/1/\infty$ queue, W_1^F represents the waiting time of a FCFS $M/G/1/\infty$ queue and $W_{c_1}^F$ represents the waiting time of a First-come, first-served (FCFS) $M/G/c_1/\infty$ queue. Here we note that the priority increases with k such that priority k represents the triage level $K+1-k$, shown in Table 1. In the following, we focus on computing W_1^F, $W_{c_1}^F$, as well as W_1^k.

Estimation of W_1^F The waiting time in a FCFS $M/G/1/\infty$ queue can be easily computed following [32]

$$W_1^F = \frac{\lambda S^2}{2(1-\rho)} \tag{9}$$

where $\rho = \lambda S$, λ represents the arrival rate of patients across all priorities, S represents the mean service time for patients across all priorities and S^2 represents the second moment of the service time of patients across all priorities.

Estimation of $W_{c_1}^F$ The waiting time in a FCFS $M/G/c_1/\infty$ queue has been estimated in [33] as

$$W_{c_1}^F \approx \left[S + \frac{S P_Q}{c_1 - \lambda S} \right] \frac{1 + C^2}{2} \tag{10}$$

where C^2 represents the SCV of the service time for patients across all priorities, λ represents the arrival rate of patients

across all priorities and S represents the mean service time for patients across all priorities. P_Q represents the probability given in [33] as

$$P_Q = \frac{(c_1\rho)^{c_1}}{c_1!(1-\rho)}\left[\sum_{t=1}^{c_1-1}\frac{(c_1\rho)^t}{t!} + \sum_{t=c_1}^{\infty}\frac{(c_1\rho)^t}{c_1!c_i^{t-c_1}}\right]^{-1} \tag{11}$$

where $\rho = \lambda S/c_1$.

Estimation of W_1^k The waiting time in a preemptive resume multi-priority $M/G/1/\infty$ queue has been well studied. Specifically, given the number of priorities K, for each priority $k = 1, \cdots, K$, the mean waiting time is expressed in [32] as

$$W_1^k = \begin{cases} \frac{(1-\rho_1)S_1+R_1}{1-\rho_1}, & \text{for } k=1 \\ \frac{(1-\rho_1-\cdots-\rho_k)S_k+R_k}{(1-\rho_1-\cdots-\rho_{k-1})(1-\rho_1-\cdots-\rho_k)}, & \text{for } k>1 \end{cases} \tag{12}$$

where $\rho_k = \lambda_k\overline{S}_{(k)}$, λ_k represents the arrival rate of patients with priority k, $\overline{S}_{(k)}$ represents the mean service time for patients with priority k (after taking into account the coupling effects), $R_k = \frac{1}{2}\sum_{i=1}^{k}\lambda_i\overline{S}^2_{(k)}$, and $\overline{S}^2_{(k)}$ is the second moment of the service time of kth-priority patients.

Estimating the mean and second moment of service time in the ED Building on Eqs. 10, 11 and 12, the estimation of W_1^F, $W_{c_1}^F$, and W_1^k are determined by the mean and second moment of service time for patients in each priority as well as the average across various priorities. Building on Eq. 1, we can attain the mean and second moment for kth-priority patients as in Eq. 13 below.

$$\overline{S}_{(k)} = S_{(k)} + P_b[mean(\min_i T_i)] \tag{13}$$

and

$$\overline{S}^2_{(k)} = S^2_{(k)} + 2S_{(k)}P_b \times mean(\min_i T_i) + P_b^2[mean(\min_i T_i)]^2$$

where $S_{(k)}$ represents the emergency service time in the ED, and $S^2_{(k)}$ represents the second moment of emergency service time in the ED.

The average service time of the first queue across various priorities, μ, equals the average of the mean service time of each priority. Namely, $\mu = 1\Big/\sum_k q_k\overline{S}_{(k)}$, where $q_k = \lambda_k/\sum_k\lambda_k$. Substituting this into Eq. 2, we can denote P_b as a function of $\sum_k\overline{S}_{(k)}$.

To compute the mean and the second moment of service time for patients in each priority, we develop a numerical algorithm, shown in Algorithm 2. The Algorithm 2 iteratively calculates $S_{(k)}$, $S^2_{(k)}$, P_b, and μ until the estimation of $S_{(k)}$ and $S^2_{(k)}$ in two consecutive iterations has a difference below a threshold. If the output is not 'No solution', then, this output can be viewed as an estimation of the exact mean and second moment of the service time within an acceptable difference. Building on this algorithm as well as Eqs. 10, 11 and 12, we can attain the estimation of the waiting time of patients to access the ED, $W_{c_1}^k$.

Algorithm 2 Numerical estimation of the mean and second moment of service time in the ED taking into account the coupling effects and the multiple priority classes

Input: The mean emergency service time for the kth-priority patients in the ED $S_{(k)}$, the second moment of emergency service time for the kth-priority patients in ED $S^2_{(k)}$, as well as the thresholds to terminate this algorithm ε_1 and ε_2

Output: The estimation of $\overline{S}_{(k)}$ and $\overline{S}^2_{(k)}$

1 Initialize $t=1$; $\varepsilon_1=\varepsilon_1^*$; $\varepsilon_2=\varepsilon_2^*$; $\overline{S}^{(0)}_{(k)} = S_{(k)}$; $\overline{S}^{2(0)}_{(k)} = S^2_{(k)}$; $\mu^{(0)} = K\Big/\sum_k\overline{S}^{(0)}_{(k)}$; $P_b^{(0)} = P_b(\mu^{(0)})$; $\overline{S}^{(1)}_{(k)} = \overline{S}_{(k)}(P_b^{(0)})$; $\overline{S}^{2(1)}_{(k)} = \overline{S}^2_{(k)}(P_b^{(0)})$; $\mu^{(1)} = K\Big/\sum_k\overline{S}^{(1)}_{(k)}$; go to step 2

2 If $\left|\overline{S}^{(t)}_{(k)} - \overline{S}^{(t-1)}_{(k)}\right| \le \varepsilon_1$ and $\left|\overline{S}^{2(t)}_{(k)} - \overline{S}^{2(t-1)}_{(k)}\right| \le \varepsilon_2$, then, $\overline{S}_{(k)} = \overline{S}^{(t)}_{(k)}$, $\overline{S}^2_{(k)} = \overline{S}^{2(t)}_{(k)}$, $\mu = \mu^{(t)}$, $P_b = P_b^{(t)}$; go to step 5; Otherwise, go to step 3

3 If $\lambda \le c_1\mu^{(t)}$, then, $R_d = \lambda$; Otherwise, $R_d = c_1\mu^{(t)}$; go to step 4

4 Calculate $P_b^{(t)} = P_b(\mu^{(t)})$ by Eq. 1, $\overline{S}^{(t+1)}_{(k)} = \overline{S}_{(k)}(P_b^{(t)})$ by Eq. 13, and $\overline{S}^{2(1)}_{(k)} = \overline{S}^2_{(k)}(P_b^{(0)})$ by Eq. 13; $t = t+1$; go to step 2

5 If $0 \le P_b \le 1$, output $S_{(k)}$ and $S^2_{(k)}$; Otherwise, output 'No solution'.

4.3 Estimation of waiting time for an ED with a fast track

While the previous sections investigated the case of ED without a fast track, we now turn our attention to the case with a fast track. A fast track is designed for patients with less serious illnesses and injuries to shorten both waiting and treatment times for these patients. More specifically, after evaluation by a triage nurse, patients with less emergent issues (triage level IV and V) are placed in the fast track [34].

In our model, a system with a fast track places paitents at triage level IV and V on an express line, and the other patients will be put on the regular line. Also c_1 beds in ED

will be split into two parts: one part for Fast Track (the number of beds is denoted as c_1^I) and one part for regular line (the number of beds is denoted as c_1^{II}). According to Eqs. (1) and (2), the averge service time (denoted as $1/\mu^I$) in the Fast Track and the average service time (denoted as $1/\mu^{II}$) in the regular line are shown as Eq. 14.

$$\begin{aligned} 1/\mu^I &= 1/\mu_1^I + P_b \times mean(\min_i T_i) \\ 1/\mu^{II} &= 1/\mu_1^{II} + P_b \times mean(\min_i T_i) \\ \mu_1^I + \mu_1^{II} &= \mu_1 \end{aligned} \tag{14}$$

where P_b is shown in Eq. 2.

The average service time (denoted as $1/\mu^F$) across all triage levels is shown in Eq. 15

$$\begin{aligned} 1/\mu^F &= \frac{1}{P_r^I \mu^I + P_r^{II} \mu^{II}} \\ P_r^I + P_r^{II} &= 1 \end{aligned} \tag{15}$$

where P_r^I is the probability of a patient whose triage level is at level I-III, while P_r^{II} is the probability of a patient whose triage level is at level IV–V.

By substituting Eq. 15 into Eq. 8, we can estimate the average waiting time of Fast Track by setting the number of beds in ED as c_1^I, while estimating the average waiting time in the regular line (acute side) by setting the number of beds in ED as c_1^{II}. Definitely, both of these waiting times are dependent on the proportion of patients who are switched over to the fast track, namely, the probability of a patient whose triage level is at level I–III P_r^I.

5 Result and discussion

In the following we investigate the necessary capacity of ED and IU in various scenarios to meet the waiting time targets from arrival to first physician assessment in the ED (shown in Table 1). For triage level I, we replace 'immediate' by '<3 min'. The other parameters in our model include the arrival rate λ as the average arrival rate of emergency patients throughout the day (shown in Table 2) and the arrival rate λ' as the average arrival rate of patients that directly access the inpatient units throughout the day (shown in Table 2). In addition, the distribution of the service time in the ED is triangular with a lower limit of 0.1 hour, upper limit of 1 hour, and mode of 0.5 hour (see Section 4.1). The distribution of the service time in the IU is triangular with a lower limit of 1 day, upper limit of 7 days, and mode of 4 days. Building on Eq. 8, we can estimate the waiting time to access the ED without a fast track as well as with a fast track, and compare the results in both scenarios in Section 5.4. The other parameters in Eqs. 1–10 can be calculated from the aforementioned parameters.

5.1 Relationship between ED and IU resource requirements

Given a set of priority specific wait time targets, the model presented here can be used to estimate the necessary capacity both in the ED and the IU in order to meet the performance targets. Figure 3 shows the impact of changes in the available IU capacity on the necessary ED capacity. In our queueing model, the average of length of stay in the IU is 4 days, the arrival rate to access the ED is 7.572 patients per hour, and we assume here that there is no fast track. Also our analytical results are verified by Monte-Carlo simulations, in which we mimic two individual queues which link with each other and calculate the necessary capacity of ED resources by repeating the simulation 50000 times.

Unsurprisingly, the necessary ED capacity increases as the size of the IU decreases. However, the impact of additional IU capacity on ED resource requirements decreases significantly as the IU increases suggesting that there is a threshold size of the IU beyond which additional increases cease to be advisable. However, attempting to reduce the size of the IU below that threshold leads to a steep increase in the required ED capacity in order to meet the same targets and therefore is likely not cost-effective.

5.2 Capacity of resources and length of stay in the IU

Clearly, the probability of congestion in the IU leading to backlogs in the ED is dependent on the rate of turnover in the IU which is regulated by the length of stay of patients. Thus, it is of interest to analyze changes in the necessary capacity of the IU/ED resources as the length of stay in the IU is varied. In the queueing model, we first vary the average length of stay in IU through 110–150 hours. Given a number of ED resources, we can determine the necessary IU

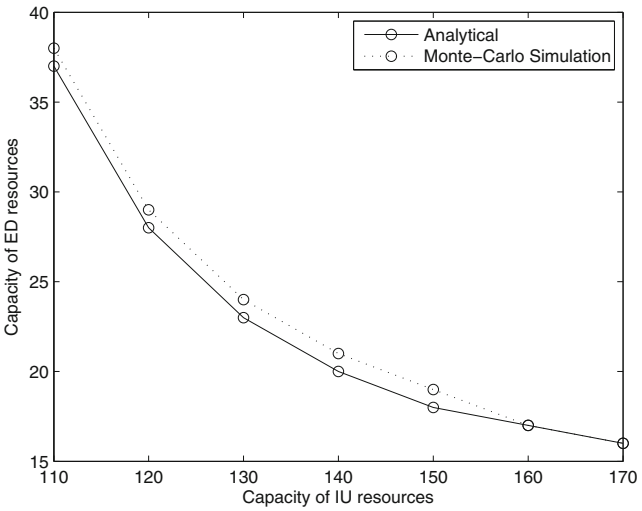

Fig. 3 Capacity of ED resources vs. capacity of IU resources

resources as the length of stay in the IU is varied. Secondly, by varying the number of ED resources through 10–20, we can obtain the 3-dimensional plot which determines how the necessary ED/IU resources change as the length of stay in the IU varies (shown in Fig. 4). In both cases, we fix the arrival rate to the ED at 7.572 patients per hour and assume that there is no fast track.

With the increase in the length of stay in the IU, the capacity of both IU resources and ED resources will, of course, increase. However, the rate of increase in the necessary IU resources as the length of stay is increased is significantly higher than for the ED (shown in Fig. 4). In other words, an increase in the length of stay in the IU causes a larger increase in the necessary IU resources. Thus, in the face of greater uncertainty in the length of stay of patients, it is preferable to carry additional or "excess" IU capacity rather than hoping to manage this problem by using the ED as a holding bay for the IU; a reality that is surprisingly common in practice.

5.3 Capacity of resources and arrival rate of ED patients

Obviously, the necessary ED and IU resources to guarantee an acceptable waiting time to see a physician are dependent on the arrival rate to the ED, so it is of interest to investigate the impact on ED and IU resource requirements of increases in this arrival rate. We focus on how much additional ED/IU capacity should be anticipated as necessary given a forecast of increasing demand. In the queueing model, we first fix the number of IU resources at 125 and determine the necessary ED resources as the arrival rate to the ED is varied. Secondly, we fix the number of ED resources at 20 and determine how the necessary IU resources change as the arrival rate to the ED is varied. In both cases, we fix the length of stay in IU at 150 hours and assume that there is no fast track.

With the increase in the arrival rate to the ED, the required capacity of ED and IU resources will, of course, increase. However, the rate of increase in the necessary IU resources as the arrival rate is increased is significantly higher than for the ED (shown in Fig. 5). In other words, an increase in the arrival rate to the ED causes a larger increase in the necessary IU resources. Thus, again in the face of greater uncertainty in the arrival rate of ED patients, it is preferable to carry additional or "excess" IU capacity rather than hoping to manage this problem by using the ED as a holding bay for the IU.

5.4 Average waiting time to access the ED with and without a fast track

The aforementioned discussion assumes that a hospital does not implement a fast track for patients at triage level IV and V. In the following, we focus on the influence of a fast track on the average waiting time to access the ED. In this scenario, the average length of stay in the IU is 4 days, and the number of IU resources is 125. In the scenario with a fast track, we allocated 20 % of ED capacity to the fast track.

As shown in Fig. 6, a fast track reduces the average waiting time across all patients to access the ED by approximately 0.3–1 hour. The explanation is as follows. The fast track provides for the prompt treatment of patients with non-life-threatening injuries. In a non-fast track system, these patients are forced to queue behind patients whose complexity requires long service times. The fast track provides short wait times for a cohort of patients who otherwise would have the longest wait, thus reducing the average wait time.

However, the reduction in the average waiting time across all patients comes at the cost of increasing the waiting time of patients who do not qualify for the fast track. As shown in

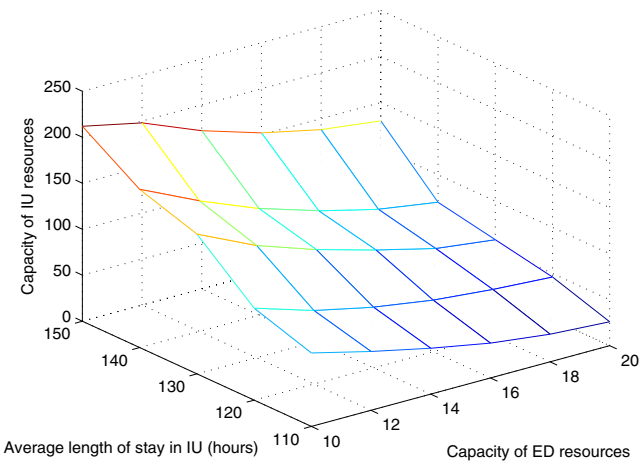

Fig. 4 Capacity of ED/IU resources vs. average length of stay in IU

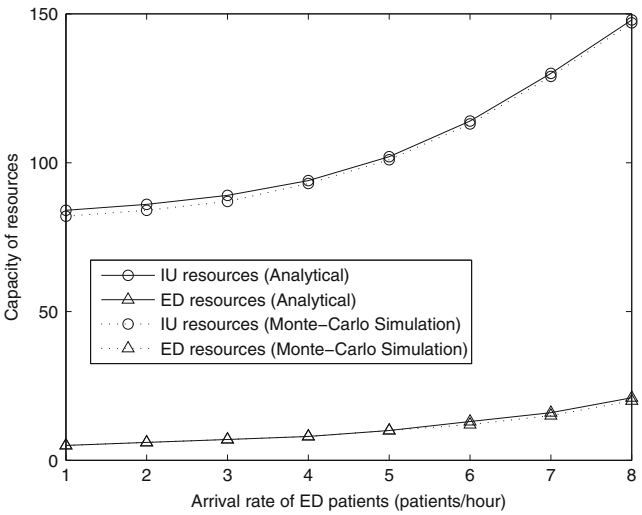

Fig. 5 Capacity of ED/IU resources vs. arrival rate of ED patients

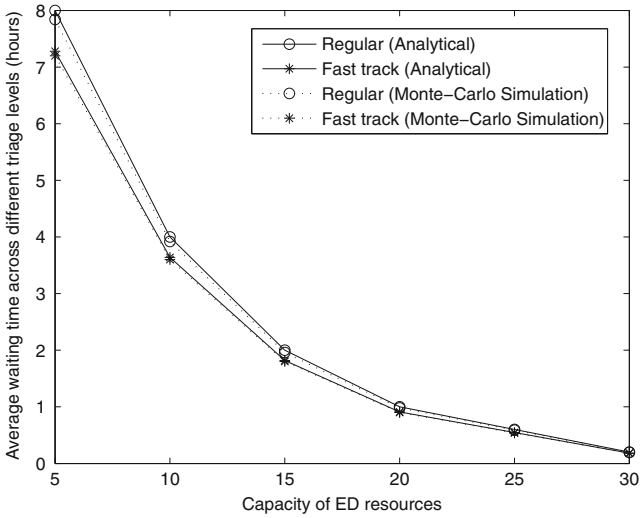

Fig. 6 Average waiting time across various triage levels vs. capaciy of ED resources

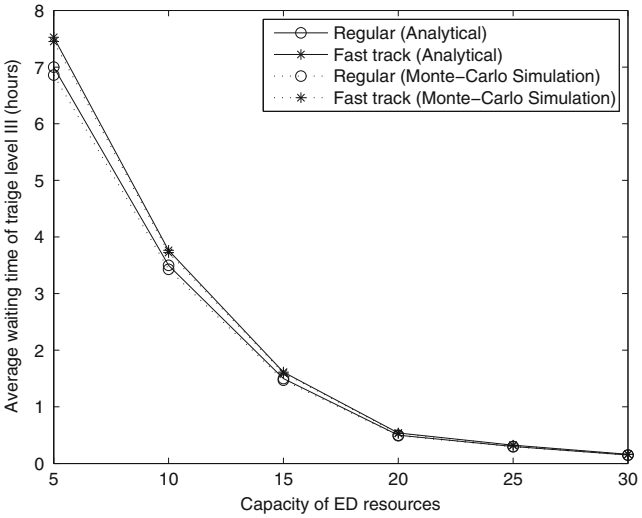

Fig. 8 Average waiting time of triage level III vs. capaciy of ED resources

Figs. 7 and 8, the average waiting time of patients at triage level III increases by around 0.1–0.5 hours in a system that implements a fast track. In this scenario, the number of ED resources is 25.

5.5 Limitations of our patient flow model

The patient flow model in our paper has the following limitations: firstly, we assume that the inpatient unit can accommodate all types of patients, but in reality, a few specialized units can only accommodate a specific type of patient. For example, an injured patient cannot stay in a cardiovascular disease unit. Secondly, in our model, we assume that one resource can only be offered to one patient, so c patients will consume c resources in the ED. However, in real scenarios, EDs with c beds can accommodate more than c patients by using hall beds and internal waiting rooms [35]. Thirdly, our model employs steady-state and average arrival rates, which do not take into account that the arrival rate is non-stationary. Finally, our model does not cover the real-life cost implications of Length of Stay (LOS) reductions.

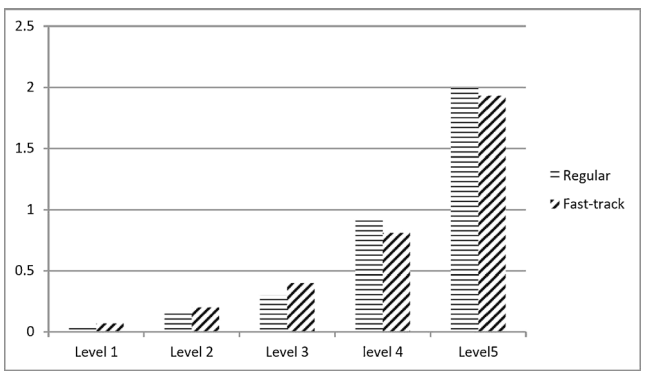

Fig. 7 Waiting time to see a doctor across various triage levels [hours]

6 Conclusions

In this paper, we develop a two-stream model to characterize the coupling effect between patient flow to access the ED and that to access the IU. Building on this patient flow model, we employ queueing theory methods to estimate the average waiting time across patients as well as the necessary ED and IU resources in order to meet target waiting times for patients at each triage level. In addition, we investigate the influence of a fast track stream on the average waiting time of patients. Our model improves on previous research by taking into account both the reality of multiple priority classes competing for ED resources and the strong potential for downstream congestion impacting on the timely access of patients to the ED.

In addition to providing hospital management with a means of determining the necessary capacity in the ED and the IU in order to meet priority specific wait time targets for timely access to the ED, this paper also provides the following insights. (1) There is a threshold size for the IU such that reductions in the IU below that threshold lead to steep increases in the necessary ED capacity in order to meet the same targets. Thus, there exists a "optimal" IU capacity such that either increasing or decreasing IU resources will lead to higher costs associated with the same performance. (2) Uncertainty in the LOS in the IU or in the arrival rate to the ED is best accommodated by carrying extra IU capacity rather than by increasing the size of the ED. (3) While a fast track can reduce the average waiting time across patients from all the triage levels, it is

accomplished by a large reduction in the wait times of patients from triage levels IV and V that offset a concurrent increase in the wait times of patients from triage level III. In other words, a fast track actually worsens the ability of an ED to provide timely access to patients whose conditions could potentially progress to a more serious problem requiring emergency intervention, such as asthma, vaginal bleeding, moderate trauma, Gastrointestinal (GI) bleeding, and acute pain. These three insights are partially verified by relevant research literature [7, 8, 36]. Erik et al. in [36] explain that increasing IU resources instead of ED resources with the aid of "IU buffers" can solve the problem of overcrowding and shows the "best buffering capacity" in IU, but this paper does not take into account the triage levels of patients. Also Cooke et al. in [7] and Miquel et al. in [8] show that the fast track can reduce the waiting time for patients who are qualified to access ED but will slightly lengthen the waiting time of the other patients. However, these two papers do not consider the coupling effects between the ED and IU.

We would like to extend our results by varying the proportion of ED resources allocated to the fast track, and develop a general rule to determine the optimal allocation depending on the proportion of demand that is eligible for the fast track and the relative resource consumption of the two groups. We would like to explore whether there exists an optimal proportion of ED resources to allocate to the fast track in order to minimize the demand of IU resources and satisfy all CTAS time requirements.

Acknowledgments Thanks to Prof. Wojtek Michalowski who worked with us and offered us quite a few fantastic ideas to finish this draft. This work was partially supported by the Natural Sciences and Engineering Research Council (NSERC) and industrial and government partners, through the Healthcare Support through Information Technology Enhancements (hSITE) Strategic Research Network, and was partially supported by Quebec MDEIE PSR-SiiRi program.

References

1. Kolb EMW, Peck J et al (2008) Reducing emergency department overcrowding - five patient buffer concepts in comparison. Simulation conference, 2008. WSC 2008. Winter
2. Akcali E, Cote MJ et al (2006) A network flow approach to optimizing hospital bed capacity decisions. Health Care Manag Sci 9(4):391–404
3. Haifeng X, Chaussalet TJ et al (2006) A model-based approach to the analysis of patterns of length of stay in institutional long-term care. IEEE Trans Inf Technol Biomed 10(3):512–518
4. Rowe BH, Bondm K, Ospina MB, Blitz S, Friesen C, Schull M, Innes G, Afilalo M, Bullard M, Campbell SG, Curry G, Holroyd B, Yoon P, Sinclair D (2006) Emergency department overcrowding in Canada: what are the issues and what can be done? [Technology overview no 21]. Ottawa, Canadian Agency for Drugs and Technologies in Health
5. Guttmann A, Schull MJ, Vermeulen MJ, Stukel TA (2011) Association between waiting times and short term mortality and hospital admission after departure from emergency department: population based cohort study from Ontario, Canada. British Medical Journal. Retrieved from http://www.bmj.com/content/342/bmj.d2983.abstract
6. Murray MJ, Bullard M, Grafstein E (2004) Revisions to the Canadian emergency department triage and acuity scale implementation guidelines. Can J Emerg Med 6:421–427
7. Cooke MW, Wilson S, Pearson S (2002) The effect of a separate stream for minor injuries on accident and emergency department waiting times. Emerg Med J 19(1):28–30. doi:10.1136/emj.19.1.28
8. Sanchez M, Smally AJ, Grant RJ, Jacobs LM (2006) Effects of a fast-track area on emergency department performance. J Emerg Med 31(1):117–120. ISSN 0736-4679, doi:10.1016/j.jemermed.2005.08.019
9. Beveridge R, Clarke B, Janes L et al (1998) Emplementation guidelines for the Canadian emergency department triage and acuity scale (CTAS). Canadian association of emergency physicians
10. Green LV, Soares J, Giglio JF, Green RA (2006) Using queuing theory to increase the effectiveness of emergency department provider staffing. Acad Emerg Med J 13(1):61–68
11. Hall RW (2006) Patient flow: the new queuing theory for healthcare, OR/MS Today
12. Au-Yeung SWM, Harrison PG, Knottenbelt WJ (2006) A queuing network model of patient flow in an accident and emergency department. Department of Computing, Imperial College of London
13. Abraham G, Byrnes GB et al (2009) Short-term forecasting of emergency inpatient flow. IEEE Trans Inf Technol Biomed 13(3):380–388
14. Lin WT (1989) Modeling and forecasting hospital patient movements: Univariate and multiple time series approaches. Int J Forecast 5(2):195–208
15. Arul E, Mark IC, Donald NG, Leo YS (2005) Using autoregressive integrated moving average (ARIMA) models to predict and monitor the number of beds occupied during a SARS outbreak in a tertiary hospital in Singapore. BMC Health Serv Res 5(36):1–8
16. Thompson S, Nunez M, Garfinkel R, Dean MD (2009) OR practice—efficient short-term allocation and reallocation of patients to floors of a hospital during demand surges. Oper Res 57(2):261–273. doi:10.1287/opre.1080.0584
17. Kolb EMW, Taesik L et al (2007) Effect of coupling between emergency department and IU on the overcrowding in emergency department. Simulation Conference, 2007 Winter
18. Rice JA (1994) Mathematical statistics and data analysis. Duxbury Resource Center
19. Patrick J, Puterman ML, Queyranne M (2008) Dynamic multipriority patient scheduling for a diagnostic resource. Oper Res 56(6):1507–1525
20. Patrick J, Puterman ML (2008) Reducing wait times through operations research: optimizing the use of surge capacity. Health Policy 3(3):75–88
21. Koizumi N, Kuno E, Smith T (2005) Modeling patients flows using a queuing network with blocking. Health Care Manag Sci 8(1):49–60
22. Whitt W (2004) A diffusion approximation for the G/GI/n/m queue. Oper Res 52(6):922–941
23. Celis M, Dennis JE, Tapia RA (1985) A trust region strategy for nonlinear equality constrained optimization. Numerical Optimization, Philadelphia: SIAM, pp 71–82

24. Gross D, Shortle JF, Thompson JM, Harris CM (2008) Fundamentals of queueing theory. Wiley, Hoboken
25. Hokstad P (1978) Approximations for the M/G/m queue. Oper Res 26:510–523
26. Miyazawa M (1986) Approximation of the queue-length distribution of an MIGI/s queue by the basic equations. J Appl Probab 23(2):443–458
27. Tijms HC, van Hoorn MH (1982) Computational methods for single-server and multi-server queues with Markovian input and general service times. Appl Probab Comput Sci 3:71–102
28. Shimshak DG, Gropp Damico D, Burden HD (1981) A priority queuing model of a hospital pharmacy unit. Eur J Oper Res 7: 350–354
29. Siddharthan K, Jones WJ, Johnson JA (1996) A priority queuing model to reduce waiting times in emergency care. Int J Health Care Qual Assur 9(5):10–16
30. Fiems D, Koole G, Nain P (2007) Waiting times of scheduled patients in the presence of emergency requests. http://www.math.vu.nl/ koole/articles/report05a/art.pdf
31. Bondi A, Buzen J (1984) The response times of priority classes under preemptive resume in M/G/m queues. In ACM Sigmetrics, pp 195–201
32. Gross D, Harris CM (1985) Fundamentals of queuing theory. Wiley, New York
33. Lee AM, Longton PA (1959) Queueing process associated with airline passenger check-in. Oper Res Q 10:56–71
34. Chan TC, Killeen JP, Kelly D et al (2005) Impact of rapid entry and accelerated care at triage on reducing emergency department wait times, length of stay, and rate of left without being seen. Ann Emerg Med 46:491–497
35. Jeffery KC, Kevin TR (2009) A multi-class queuing network analysis methodology for improving hospital emergency department performance. Comput Oper Res 36(5):1497–1512
36. Kolb EMW, Schoening S, Peck J, Lee T (2008) Reducing emergency department overcrowding: five patient buffer concepts in comparison. In: Mason S, Hill R, Mnch L, Rose O (eds) Proceedings of the 40th conference on winter simulation (WSC '08). Winter Simulation Conference, pp 1516–1525